Explorer des corpus à l'aide de CasSys. Application au Corpus d'Orléans

Denis Maurel, Nathalie Friburger, Iris Eshkol, Jean-Yves Antoine

Explorer des corpus à l'aide de CasSys. Application au *Corpus d'Orléans*

Denis MAUREL[1], Nathalie FRIBURGER[1], Iris ESHKOL[2], Jean-Yves ANTOINE[1]

[1]Université François Rabelais Tours, laboratoire d'informatique
[2]Université d'Orléans, laboratoire ligérien de linguistique

1 Introduction

Cet article présente un outil d'exploration de corpus, *CasSys*, facilement paramétrisable par les linguistes, permettant de reconnaître des motifs même complexes et de les baliser, éventuellement par des balises XML. Ce balisage automatique peut ensuite être révisé par un expert. CasSys est donc un outil d'exploration de corpus, mais également d'annotation enrichie semi-supervisée.

Deux exemples réels complèteront cette présentation : la recherche des entités nommées du *Corpus d'Orléans* et l'utilisation de ces entités pour connaître des informations sur les personnes répondant à l'enquête qui constitue ce corpus. Ce travail a bénéficié du financement du projet ANR *Variling* et d'un projet Feder Région Centre. Il a aussi été testé dans le cadre de l'évaluation *Ester2* (*campagne d'évaluation des systèmes de transcription enrichie d'émissions radiophoniques*)[1].

2 Le système CasSys

Le système CasSys [Friburger, 2002] [Friburger, Maurel, 2004] est un programme permettant le passage sur un corpus de transducteurs "en cascade" (c'est-à-dire les uns après les autres) dans un ordre défini [Abney, 1996]. Ces transducteurs sont des graphes Unitex [Paumier, 2003], facilement manipulables et modifiables par un linguiste grâce à une interface conviviale. Il est donc possible de définir des cascades pour différents balisages.

1. Annotation de corpus : moyennant quelques adaptations, CasSys peut s'adapter à tout corpus, nous l'avons appliqué au journal Le Monde, à Eslo1 et Ester2, pour la reconnaissance des entités nommées (cascade *CasEN*).

2. Exploration de corpus : par exemple, une fois les entités nommées balisée, la cascade *CasDen* reconnaît les entités dénommantes, c'est-à-dire les informations concernant la personne interrogée (composition familiale, profession, lien avec Orléans…).

La Figure 1 présente un exemple de transducteur de la cascade CasEN. Celui-ci reconnaît et balise les établissements de soin. Des transducteurs précédents ont déjà reconnu une personne (idxPerson), éventuellement avec un titre (idxTitre), ou une organisation (idxOrg), ou encore

[1] http://www.afcp-parole.org/ester

un lieu (idxLoc)… Par exemple, *hôpital Jean Rostand*, *clinique Docteur Calabet d'Agen*, etc. qui seraient placés entre les balises <ENT type=loc.fac> et </ENT>.

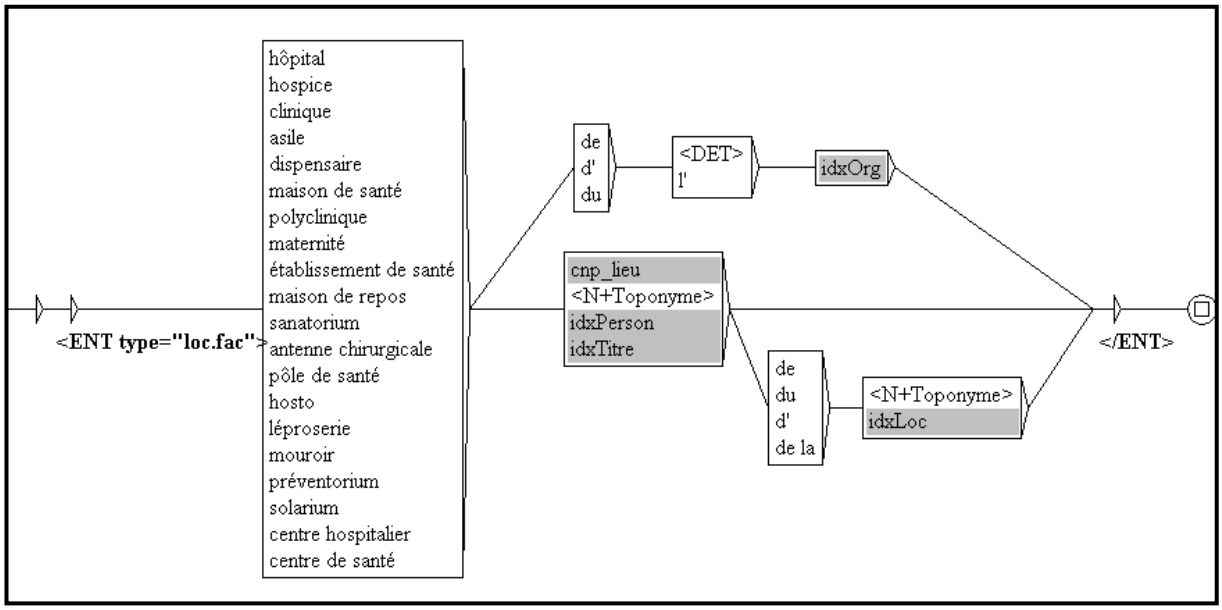

Figure 1 : un exemple de transducteur

3 Le *Corpus d'Orléans*

Le corpus sur lequel nous avons travaillé est l'*Enquête sociolinguistique à Orléans* (*Eslo1*), plus précisément, les cent vingt premières heures d'interview transcrites (cent cinq fichiers *Transcriber*, soit 31 004 Ko). Nous comptons utiliser aussi bientôt le corpus *Eslo2*, en cours de constitution, en partant des acquis d'Eslo1, une nouvelle enquête a été mise en chantier par le LLL. Il s'agit, à quarante années de distance, de constituer un corpus comparable dans le produit attendu et dans les modalités de la collecte (400 heures, environ 6 000 000 de mots).

Les principales conventions de transcription sont l'absence de ponctuation et de majuscule en début d'énoncé ainsi qu'une transcription orthographique normée (majuscule pour les noms propres, transcription des chiffres et des dates en toutes lettres avec les traits d'union si nécessaire, termes épelés notés entièrement en majuscule). Le questionnaire de l'entretien contient tout d'abord des questions préliminaires (Depuis combien de temps habitez-vous Orléans ?, Qu'est-ce qui vous a amené à vivre à Orléans ?, Est-ce que vous vous plaisez à Orléans ?, etc.), puis des questions sur le travail et les loisirs du locuteur et des membres de sa famille, ce qui explique la présence d'un nombre important d'entités nommées dans le corpus.

4 La cascade CasEN

La cascade CasEN a tout d'abord été conçue pour reconnaître les entités nommées de corpus journalistiques (essentiellement *Le Monde*) [Friburger, 2002]. En 2006, dans le cadre du projet ANR Variling, elle a été reprise et adaptée au corpus d'Orléans. De plus, le balisage des entités nommées a été modifié pour correspondre à celui de la campagne Ester 2, à quelques ajouts près. Cette typologie est présentée Figure 2.

personne (*pers*)	humain réel ou fictif (*pers.hum*) animal réel ou fictif (*pers.anim*)	
fonction (*fonc*)	politique (*fonc.pol*) militaire (*fonc.mil*) administrative (*fonc.admi*) religieuse (*fonc.rel*) aristocratique (*fonc.ari*)	
organisation (*org*)	politique (*org.pol*) éducative (*org.edu*) commerciale (*org.com*) non commerciale (*org.non-profit*) média & divertissement (*org.div*) géo-socio-administrative (*org.gsp*)	
lieu (*loc*)	géographique naturel (*loc.geo*) région administrative (*loc.admi*) axe de circulation (*loc.line*) construction humaine (*loc.fac*)	
	adresse (*loc.addr*)	adresse postale (*loc.addr.post*) téléphone et fax (*loc.addr.tel*) adresse électronique (*loc.addr.elec*)
production humaine (*prod*)	moyen de transport (*prod.vehicule*) récompense (*prod.award*) œuvre artistique (*prod.art*) production documentaire (*prod.doc*)	
date et heure (*time*)	date (*time.date*)	date absolue (*time.date.abs*) date relative (*time.date.rel*)
	heure (*time.hour*)	
montant (*amount*)	valeur physique (*amount.phy*)	âge (*amount.phy.age*) durée (*amount.phy.dur*) température (*amount. phy.temp*) longueur (*amount. phy.len*) surface et aire (*amount.phy.area*) volume (*amount.phy.vol*) poids (*amount. phy.wei*) vitesse (*amount. phy.spd*) autre (*amount.phy.other*)
	valeur monétaire (*amount.cur*)	

Figure 2 : La typologie Ester[2]

Cependant ces annotations ne nous ont pas paru suffisantes sur deux points. Tout d'abord, il nous a semblé intéressant d'annoter et de détailler l'ensemble des informations sur les personnes, lorsque celles-ci constituent une expansion classifiante de l'entité. Par exemple simplement sa civilité, mais aussi son origine géographique (gentilé, ethnique…), sa

[2] http://www.afcp-parole.org/ester/docs/Conventions_EN_ESTER2_v01.pdf

profession, sa religion, son appartenance politique, etc. De plus nous avons aussi ajouté un balisage des dynasties. Ensuite, nous avons aussi complété les différents types d'Ester par un typage des évènements, parmi lesquels nous avons distingué les faits historiques et les différentes manifestations sportives, culturelles, etc. La Figure 3 détaille ces ajouts.

pers.hum.tit	les civilités (M. Mme, Melle, Monsieur, Madame…)
pers.hum.gent	les gentilés (Tourangeau, Parisien…) les adjectifs toponymiques (tourangeau, parisien…)
pers.hum.occ	les professions (avocat, journaliste…)
pers.hum.sp	les sports (coureur cycliste, footballeur…)
pers.hum.art	les artistes (violoniste, sculteur…)
pers.hum.nat	la nationalité (français, hollandais…)
pers.hum.rel	la religion (musulman, catholique…)
pers.hum.pol	la politique (communiste, socialiste…)
pers.hum.fonc	les titres professionnels (Professeur, Maitre, Docteur…)
pers.hum.dynasty	(les Bourbons, les Windsor…)
event	les évènements
event.hist	l'histoire (la Seconde Guerre mondiale, la Prise de la Bastille…)
event.manif	les manifestations sportives, artistiques (les Jeux olympiques, les Francofolies…)

Figure 3 : Complément aux annotations d'Ester

La cascade *casEN* commence par lancer le programme de préanalyse d'*Unitex*, en utilisant un graphe inspiré de celui de [Dister, 2007], puis le dictionnaire Delas [Courtois., Silberztein, 1990] et des dictionnaires spécifiques contenant 28 341 prénoms, 31 580 professions [Gazeau, Maurel, 2006], 3 016 sigles, 114 511 noms propres (et dérivés de noms propres) extraits de Prolexbase [Tran, Maurel, 2006], 497 noms d'animaux, 296 noms de sports, 110 noms de monnaies, 53 noms de marques de voiture et 26 noms de quotidiens. Elle est suivie de 152 transducteurs passés en cascade.

Voici quelques exemples de balisage :

- il y a deux ans une euh <ENT type="pers.hum.nat"> anglaise </ENT>
- chez moi <ENT type="pers.hum"> Bérénice Nutal </ENT>
- dans les <ENT type="org.com"> PTT </ENT>
- moi je suis native de <ENT type="loc.admi"> Pithiviers </ENT> j'aime mieux <ENT type="loc.admi"> Orléans </ENT>
- <ENT type="pers.hum"> Sophie </ENT> viens voir

- oh j'ai une <ENT typr="prod.art"> encyclopédie Quillé </ENT> j'ai le
- <ENT type="time.date.abs"> en dix-neuf cent trente-huit </ENT>
- je crois que le <ENT type="org.pol"> ministère de l'Education National </ENT>
- le <ENT type="org.edu"> lycée Pothier </ENT> et les élèves qui vont au <ENT type="org.edu"> lycée Benjamin Franklin </ENT>
- euh passer quelques jours sur la <ENT type="loc.geo"> Côtes d'Azur </ENT>
- euh je suis je travaille à l'<ENT type="loc.fac"> hôpital d'Orléans </ENT> quoi
- en <ENT type="loc.admi"> Norman- Normandie </ENT> peut-être ?
- parce que nous avions un <ENT type="loc.fac"> magasin Phildar </ENT> juste en face de chez nous
- oui c'est un petit <ENT typr="prod.art"> dictionnaire Larousse </ENT>
- qui vont se charger au maximum quand le gars aura le dos tourné vous connaissez ben l'histoire de <ENT type="pers.hum"> Marius et Fanny </ENT> et tout ça
- j'ai des copains qui y travaillent et c'est très intéressant ils ont fait le <ENT type="loc.fac"> pont de Tancarville </ENT> euh
- dans la <ENT type="loc.line"> rue Royal</ENT> euh

L'ensemble du corpus Eslo1 a ainsi été annoté et ces annotations ont été vérifiées manuellement ensuite. L'évaluation des résultats du passage de la cascade CasEN a été détaillée dans [Maurel et al., 2009]. Un résumé de cette évaluation est présenté Figure 4. Puisqu'une relecture des entités trouvées était prévue dans le projet, il nous a paru intéressant de mesurer tout d'abord la simple détection des entités, en acceptant d'éventuelles erreurs de typage ou de bornage. CasEN a aussi été évaluée dans le cadre de la campagne Ester2 [Galliano et al., 2009].

Entités	Non typée	Partielle	Complète
Rappel	94,0%	88,4%	87,5%
Précision	97,8%	92,0%	91,1%

Figure 4 : Résultats de l'évaluation

5 La cascade CasDen

Le corpus Eslo1 était, à l'origine, une enquête sociolinguistique. Il nous a paru intéressant, dans le cadre du projet Variling, de proposer au lecteur des informations sur les personnes interviewées et sur leur famille. Nous avons appelé ces donnés *entités dénommantes*. Elles permettent de mieux connaître sociologiquement le locuteur. Peut-être certaines de ces

informations seront d'ailleurs cachées dans le cadre de la distribution libre du corpus [Eshkol, 2007], afin de respecter l'anonymat des témoins [Baude, 2006]. Nous avons donc créé une nouvelle cascade, *CasDen*, pour justement repérer automatiquement ce genre d'information. Celle-ci passe non pas sur le texte originel, mais sur le texte balisé par la cascade CasEN.

Nous avons donc défini de nouveaux types pour décrire la personne qui parle ou dont on parle (identité, famille, travail, syndicat, âge, origine…). Ces types sont présentés Figure 5. Parfois nous avons ajouté des informations quantitatives (nombre d'enfants, âge…) dans les balises.

pers.speaker	la personne interviewée
pers.spouse	son époux ou épouse
pers.child	ses enfants
pers.parent	ses autres liens de parenté
identity.age	l'âge
identity.origin	l'origine géographique
identity.birth	la date de naissance
identity.arrival	la date d'arrivée à Orléans
identity.children	l'identité de ses enfants
work.occupation	le métier
work.field	le domaine professionnel
work.location	le lieu de travail
work.business	l'entreprise
trade union	l'appartenance syndicale

Figure 5 : Typologie des annotations de la cascade *CasDen*

Voici par exemple le traitement d'une question sur l'arrivée de l'interviewé à Orléans :

- depuis combien de temps habitez vous <ENT type="loc.admi">Orléans</ENT> ?
  <DE type="pers.speaker"><DE type="identity.origin">
  <Turn speaker="spk1" startTime="6.754" endTime="10.88">
  oh ça fait <ENT type="time.date.rel">neuf ans</ENT> depuis dix neuf cent soixante</DE></DE>

Ou encore une question sur son travail :

- et qu'est ce que vous faites comme travail ?
  <Turn speaker="spk1" startTime="40.394" endTime="43.041">

<DE type="pers.speaker">je suis<DE type="work.occupation"> contrôleur divisionnaire<DE type="work.occupation"> au <ENT type="org.com"> PTT </ENT></DE></DE></DE>

Ou son syndicat :

- quel est votre syndicat ?
  <Turn speaker="spk1" startTime="1152.961" endTime="1159.44">
  <DE type="syndicat"> <ENT type="org">Force Ouvrière</ENT></DE>

Certaines indications ne découlent pas d'une question, mais se trouvent disséminées ici ou là dans la conversation. Par exemple, une question sur la connaissance du patrimoine local va introduire une indication sur la date d'arrivée à Orléans :

- mais est-ce qu'il y a quelque chose dans la région ou surtout à<ENT type="loc.admi"> Orléans</ENT> que vous pouvez recommander ?
  <Turn speaker="spk1" startTime="1614.656" endTime="1629.854">
  oui hein ça ça dépend des goûts des personnes aussi hein vous avez des des personnes qui
  …
  à ce moment là je peux lui expliquer je peux toujours lui dire quand même <DE type="pers.speaker"><DE type="identity.arrival"><ENT type="time.date.rel">depuis neuf ans</ENT> que je suis là</DE></DE> je commence à connaître la ville

Et la description de la recette de l'omelette peut être l'occasion de glisser son origine géographique :

- comment qu'on fait une omelette ?
  …
  on bat tout ensemble
  euh
  on met dans un peu d'eau je crois
  on mélange un peu d'eau
  enfin on assaisonne sel poivre euh
  <DE type="pers.speaker">nous en <DE type="identity.origin"><ENT type="loc.admi"> Lorraine</ENT></DE></DE> on
  on découpe des petits des petits morceaux de lards qu'on fait frire avant

Une évaluation de cette cascade est aussi présentée dans [Maurel et al., 2009]. La précision est de 94,2% et le rappel de 84,4%.

6 Conclusion

Nous avons présenté le système de cascade de transducteur CasSys qui permet une annotation des textes par un balisage. Comme CasSys utilise des graphes réalisés à l'aide de la plateforme linguistique Unitex, l'utilisateur bénéficie d'une interface graphique très conviviale pour leur réalisation.

Nous avons montré que CasSys est un système adaptable à différents types de corpus et différentes problématiques nécessitant un balisage, que ce soit de l'annotation ou de l'exploration de corpus. Signalons qu'en plus de la détection d'entités nommées et d'entités dénommantes, le système CasSys a aussi été utilisé dans le projet ANR Epac pour découper le corpus en segments syntaxiques minimaux (*chunks*) [Antoine et al., 2008].

Grâce au soutien de la Région Centre (projet Feder en cours), CasSys devrait être intégré à la plateforme Unitex d'ici la fin 2010.

Références

ABNEY S. (1996), Partial Parsing via Finite-State Cascades, *Workshop on Robust Parsing*, 8th *European Summer School in Logic, Language and Information*, Prague, Tchèquie, 8-15.

ANTOINE J. Y., MOKRANE A., FRIBURGER N. (2008), Automatic Rich Annotation of Large Corpus of Conversational transcribed speech: the Chunking Task of the EPAC Project, Sixth language resources and evaluation conference (LREC 2008), Marrakech, Maroc, 28-30 mai.

BAUDE O. (2006), Corpus oraux. Guide des bonnes pratiques, Presses universitaires d'Orléans.

COURTOIS B., SILBERZTEIN M. (1990), Dictionnaires électroniques du français, *Langues française*, 87:11-22.

DISTER A. (2007). *De la transcription à l'étiquetage morphosyntaxique. Le cas de la banque de données textuelles orales VALIBEL*. Thèse de linguistique. Université catholique de Louvain.

ESHKOL I. (2007), Entrer dans l'anonymat. Etude des entités dénommantes dans un corpus oral. Actes du colloque NOMINA2007. Ed. Narr, Tübingen (à paraître).

FRIBURGER N. (2002), *Reconnaissance automatique des noms propres ; application à la classification automatique de textes journalistiques*, Thèse de doctorat d'informatique, Université François Rabelais Tours.

FRIBURGER N., MAUREL D. (2004), Finite-state transducer cascades to extract named entities in texts, *Theoretical Computer Science*, vol. 313, 94-104.

GALLIANO S., GRAVIER G., CHAUBARD L. (2009), The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts, *Interspeech 2009*.

GAZEAU M. A., MAUREL D. (2006), Un dictionnaire INTEX de noms de professions : quels féminins possibles ?, *Cahiers de la MSH Ledoux*, 115-127.

MAUREL D., FRIBURGER N., ESHKOL I. (2009), Who are you, you who speak? Transducer cascades for information retrieval, *4th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, Poznań, Poland, 220-223.

PAUMIER S. (2003), *De la Reconnaissance de Formes Linguistiques à l'Analyse Syntaxique*, Thèse de Doctorat en Informatique, Université de Marne-la-Vallée.